Mind the Cultural Gap: Bridging Language-Specific DBpedia Chapters for Question Answering

Elena Cabrio, Julien Cojan, Fabien Gandon

Mind the cultural gap: bridging language specific DBpedia chapters for Question Answering

Elena Cabrio, Julien Cojan and Fabien Gandon

Abstract In order to publish information extracted from language specific pages of Wikipedia in a structured way, the Semantic Web community has started an effort of internationalization of DBpedia. Language specific DBpedia chapters can contain very different information from one language to another, in particular they provide more details on certain topics, or fill information gaps. Language specific DBpedia chapters are well connected through instance interlinking, extracted from Wikipedia. An alignment between properties is also carried out by DBpedia contributors as a mapping from the terms in Wikipedia to a common ontology, enabling the exploitation of information coming from language specific DBpedia chapters. However, the mapping process is currently incomplete, it is time-consuming as it is performed manually, and it may lead to the introduction of redundant terms in the ontology. In this chapter we first propose an approach to automatically extend the existing alignments, and we then present an extension of QAKiS, a system for Question Answering over Linked Data that allows to query language specific DBpedia chapters relying on the above mentioned property alignment. In the current version of QAKiS, English, French and German DBpedia chapters are queried using a natural language interface.

Key words: Question Answering, Linked Data, DBpedia, multilingualism

1 Introduction

The Semantic Web provides a framework to transform the access to information by adding machine-readable linked data and the semantics of their schema to the

Elena Cabrio
INRIA Sophia Antipolis, and EURECOM, France, e-mail: elena.cabrio@inria.fr

Julien Cojan and Fabien Gandon
INRIA Sophia Antipolis, France, e-mail: firstname.lastname@inria.fr

human-readable textual content, to facilitate automated processing and integration of the vast amount of available information on the web. The Semantic Web is an extension of the classical web, and the data and schemas it adds, co-exist with the documentary representations, that were already available and linked on the web. Moreover, more and more web sites are adding direct access to the data they use to generate their pages and enhance existing services they offer by semantic data. This not only allows interoperability, reusability and potentially unforeseen applications of opened data, but it leads to a unique situation in which large amounts of information is available, both in textual form for human consumption, as well as in structured form in line with standard shared vocabularies for consumption by machines.

A very important case of such web sites offering strongly tied texts and data, is the couple Wikipedia-DBpedia. Collaboratively constructed resources, such as Wikipedia, have grown into central knowledge sources providing a vast amount of updated information accessible on the Web, essentially as pages for human consumption. From such corpora, structured information has been extracted and stored into knowledge bases - e.g. the DBpedia project, Bizer et al. (2009) - that cover a wide range of different domains and connect entities across them. The original DBpedia project has then been mirrored at other sites for the Wikipedia content in other languages than English: we refer to the collection of such DBpedia projects as "language specific DBpedia chapters".[1] Language specific DBpedia chapters are well connected through instance interlinking, extracted from Wikipedia (more details are provided in Section 2). An alignment between properties is also carried out by DBpedia contributors as a mapping from the terms used in Wikipedia to a common ontology, enabling the exploitation of information coming from the language specific DBpedia chapters. At the same time, language specific DBpedia chapters can contain different information from one language to another, providing more specificity on certain topics, or filling information gaps. For instance, when looking for the nationality of Barack Obama on the English DBpedia chapter, we notice that there is no property *nationality* directly linking Obama to the United States. Such information can instead be found in the French DBpedia chapter, the second biggest chapter. Moreover, the knowledge of certain instances and the conceptualization of certain relations can be biased according to different cultures, and this is reflected in the structure and content of such collaboratively constructed resources. No information is provided in English Wikipedia and DBpedia, for instance, for the French musical group "Les Frères Jacques", or for the French writer Jean-Bernard Pouy.

Being able to exploit all the amount of multilingual information would bring several advantages to systems that harvest information from Wikipedia and DBpedia - and, more generally, from the Multilingual Semantic Web (Buitelaar, Choi, Cimiano, & Hovy, 2013) - automatically, both considering *i)* the intersection of such resources in different languages to detect contradictions or divergences, and *ii)* the union of such resources, to fill information gaps (cross-fertilization among languages). Also Rinser, Lange, and Naumann (2013) highlight the importance of

[1] `http://wiki.dbpedia.org/Internationalization/Chapters`

mapping the attributes of the infoboxes across different language versions, to increase the information quality and quantity in Wikipedia.

In the context of Natural Language (NL) Question Answering (QA) over Linked Data, a system which is able to exploit information coming from the multilingual and parallel versions of DBpedia would increase its probability to retrieve a correct answer (i.e., its recall). Given the multilingual scenario, attributes are labeled in different natural languages. The common ontology enables to query the multiple DBpedia chapters with the same vocabulary on the mapped data. Unfortunately, the cross-language mapping process of properties among language specific DBpedia chapters is currently incomplete, it is time consuming since it is performed manually, and it may lead to the introduction of redundant terms in the ontology, as it becomes difficult to navigate through the existing vocabulary. Moreover, several problems arise concerning both the variety and ambiguity of properties extracted from Wikipedia Infoboxes (e.g. attributes names are not always sound, often cryptic or abbreviated), and the fact that they are specific to a particular language.
In this chapter, we tackle the following research question:

How to fill the gaps between language specific DBpedia chapters for QA?

Given the complexity of our research question, in this chapter we narrow its scope, answering to the following subquestions:

(1) *how to benefit from querying language specific DBpedia data sets in the current mapping progress?*
(2) *how to safely extend the property alignments?*
(3) *how can QA systems benefit from querying language specific DBpedia chapters?*

In this chapter, we do not make use of general alignment techniques, and we do not enter in the merits of the related discussions.[2] We rather exploit the existing manually created alignments.

In the first part of the chapter we carry out a comparative analysis of property alignment in language specific DBpedia chapters, considering English and French DBpedia chapters as a case study, and highlighting the current status of the property alignment between them. Moreover, we propose an approach to automatically extend the existing alignments taking advantage of Wikipedia and DBpedia structures.

In the second part of the chapter, we present an extension of QAKiS (Cabrio et al., 2012), a system for Question Answering over Linked Data, that allows to query language specific DBpedia chapters exploiting the above mentioned property alignments. Extending QAKiS to query language specific data sets goes in the direction of enhancing users consumption of semantic data originally produced for a different culture and language, overcoming language barriers.[3]

The reminder of the chapter is structured as follows. Section 2 provides an analysis of the current status of property alignments in language specific DBpedia chapters (focusing on the English and French versions), while Section 3 proposes an

[2] For an overview, see Trojahn et al. [this volume]

[3] Currently a hot topic, see the Multilingual Question Answering over Linked Data challenge (QALD-3 and 4) `http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/index.php?x=home`

approach to extend the current mappings. Section 4 describes QAKiS extension to query language specific DBpedia chapters. Section 5 discusses the related work in the literature; conclusions end the chapter.

2 DBpedia property alignment current status

As introduced before, DBpedia (Bizer et al., 2009) is a community effort to extract structured data from Wikipedia, and to publish it as linked data. At the beginning, it only contained data extracted from the English Wikipedia, while in the most recent period, efforts to integrate data extracted from chapters of languages different from English have arisen (e.g. for German, Spanish, French and Italian). However, in the current state of affairs, the content is still focused on the English chapter, due to the fact that naming conventions limit the coverage of other chapters, and the fact that English is the biggest chapter.

Language specific DBpedia chapters have been created following the Wikipedia structure (Kontokostas et al., 2012): each chapter contains therefore data extracted from Wikipedia in the corresponding language, and so reflects local specificity. Data are published in RDF, and are structured in triples `<subject, predicate, object>` where the *subject* is an instance corresponding to a Wikipedia page, the *predicate* is a property from the DBpedia ontology or from other vocabularies (e.g. foaf, dublin core, georss), and the *object* is either a literal value or another instance.

Data from different DBpedia chapters are connected by several alignments: *i) instances* are aligned according to the inter-language links that are created by Wikipedia editors to relate articles about the same topic in different languages. As shown in Rinser et al. (2013) these correspondences are far from being perfect, but a simple filter applied before data publication in DBpedia significantly improves its quality; *ii) properties* mostly come from template attributes, i.e. structured elements that can be included in Wikipedia pages to display structured information, the most common being the infoboxes. The generic template extraction that creates property URIs from their textual names has the inconvenient of generating a large variety of properties, as well as ambiguous terms. For instance, both properties `propEn:birthDate`[4] and `propEn:dateOfBirth` appear in English DBpedia with the same meaning. On the contrary, the property `propEn:start` is used to indicate both the starting place of a route (e.g. the first station on a railway line), and the date of the beginning of an event. Moreover, as introduced before, the terms used for properties are language-dependent.

To overcome these limitations, a common ontology and mappings from template definitions to the ontology vocabulary are being collaboratively edited by the DBpedia community.[5] For instance, the attributes *date of birth* and *birth date* are mapped to the ontology property `dbo:birthDate`[4] in the description of a person, and

[4] For simplification, we use here the shorthand `propEn:` for `http://en.dbpedia.org/property/` and `dbo:` for `http://dbpedia.org/ontology/`

[5] On the wiki `http://mappings.dbpedia.org`

the attribute *start* is mapped to `dbo:routeStart` when describing a road, to `dbo:startDate` when describing an event. This term normalization effort has the goal to improve the alignment of properties among language specific DBpedia chapters. It is, however, ongoing work, and needs constant maintenance as Wikipedia templates evolve over time. Assistance tools for mapping editions, as well as automated techniques to extend the resulting alignments are becoming therefore important issues to address.

As a case study to analyze the current state of affairs of property alignment in language specific DBpedia chapters, we consider the datasets of English and French DBpedia. While the English chapter is the biggest and the most complete, with about 400 million triples[1] and 345 templates mapped, the French chapter is the second chapter in size (∼130 million triples, and 42 templates mapped). In our analysis, for each object property `prop` we compare the triples <`subject`, `prop`, `object`> from English and French DBpedia on aligned pairs of instances `subject` and `object`. That is, triples <$\texttt{subject}_{\texttt{fr}}$, `prop`, $\texttt{object}_{\texttt{fr}}$> from French DBpedia are transposed into <$\texttt{subject}_{\texttt{en}}$, `prop`, $\texttt{object}_{\texttt{en}}$>, where $\texttt{subject}_{\texttt{en}}$ and $\texttt{object}_{\texttt{en}}$ are respectively instances of English DBpedia related to $\texttt{subject}_{\texttt{fr}}$ and $\texttt{object}_{\texttt{fr}}$ through the relation `owl:sameAs`. These triples are compared with triples <`subject`, `prop`, `object`> from English DBpedia such that `subject` and `object` are also related to French instances with relation `owl:sameAs`.

Figure 1 describes the possible outcomes of such a comparison. In case *a)* we have the same value for the property in both the English and the French chapters. For instance, for the subject *Barack Obama* the property `birthPlace` is present in both the English and the French versions, with the same value. In this case, the French chapter does not bring new information, except a confirmation of values found in the English chapter. In *b)* we also have the same property in both English and French chapters, but this time with different values. In *c)* we have values for the property in the English chapter only: in the example of *Barack Obama*, the property `residence` is present for the English chapter (with the value *White House*), while it is missing in the French version. In *d)* we have a value for the property in the French chapter only, again in the example of *Barack Obama*, the property `nationality` is missing for the English DBpedia chapter, while it is present in the French version (with the value *États-Unis*, i.e. *United States*).

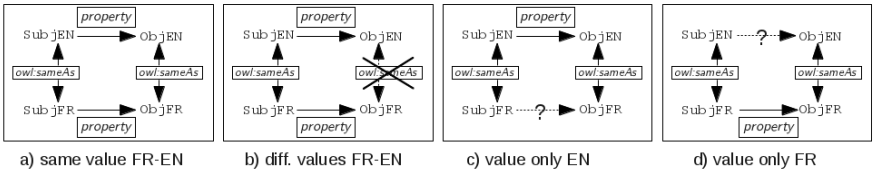

Fig. 1 Outcomes of the comparison between EN and FR chapters

There can be two reasons for differing values in case *b)* (Fig. 1): *i)* a disagreement between the two datasets, produced either by an error in one of them, or reflecting a different viewpoint (e.g. for properties of type `owl:functionalProperty`); or *ii)* the values reported in the two chapters are complementary, often providing a different granularity level (e.g. city vs country for the birth place of *Henry Lawson*). The first case can be interestingly exploited to automatically detect inconsistencies among the data which can help the Wikipedia community to improve information quality across language versions. The second one brings additional information on the subject, but it could also help to infer relationships between the values (for instance that the city where *Henry Lawson* was born is in his country of birth).

The same comparison has been carried out for datatype properties over triples `<subject, prop, val>` with aligned instances `subject`. For every property `prop`, we count: *a)* how many `subject` have the same values with `prop` in French and English, *b)* how many have at least one different value, and how many have only values either *c)* in the English or *d)* in the French DBpedia. We observed that the ratio between the number of values that are the same in English and French chapters and the number of values that are different is lower for datatype properties than for object properties. This is true in particular for string literals, as most of them are expressed in their respective chapter language (we did not compare neither instance labels nor abstracts). Nevertheless, we kept these properties in our comparison, as some of them bring information that can be exploited in a different language, for instance for people's names.

Reflecting the different progression of the mapping task between French and English DBpedia, 217 ontology properties are currently used in French DBpedia, compared to more than 1000 in English DBpedia.

Table 1 shows some statistics resulting from the comparison between English and French DBpedia. In particular, it shows some of the (object) properties for which French DBpedia presents the highest number of values not present in the English version, i.e. the properties to which the French chapter can contribute most. Moreover, it provides the total number of pairs (subject, property) that *a)* have a value in common in English and French chapters, *b)* have different values in the two chapters, *c)* have only values in English chapter, *d)* have only values in French chapter. Two intermediate sums are also given for the object properties and for the datatype properties. These sums show overall that the aligned data from the French and English chapters are quite complementary. About 47% of the data from the French DBpedia expressed in the common ontology cannot be found in English DBpedia (column *d* vs. *a*+*b*+*d*), and about 80% of the data from the English DBpedia expressed in the common ontology cannot be found in French DBpedia (column *c* vs. *a*+*b*+*c*). The values provided in Table 1 for the column *d) "only FR value"* confirm our initial intuition that being able to exploit language specific DBpedia chapters provides an additional amount of information both specific to a certain culture (for instance, concerning French habits, food or minor musical groups), and to fill information gaps (for instance, missing links in the English chapters).

	a) same value FR-EN (%)	b) diff. values FR-EN (%)	c) value only EN (%)	d) value only FR (%)
dbo:nationality	1536 (3.8)	437 (1)	11825 (29.6)	26074 (65.6)
dbo:birthPlace	14139 (17.4)	1965 (2.5)	49754 (61.3)	15279 (18.8)
dbo:region	22178 (44.5)	676 (1.4)	14397 (29)	12502 (25.1)
total object properties	239321 (14.6)	40232 (2.6)	1046532 (64.3)	305452 (18.7)
total datatype properties	104262 (7.6)	134995 (9.8)	976025 (71.2)	155134 (11.4)
total	343583 (11.4)	175227 (5.8)	2022557 (67.3)	460586 (15.5)

Table 1 Statistics resulting from the comparison of the FR and EN DBpedia chapters. For every property `prop`, column *a)* shows how many `subject` have the same values with `prop` in French and English, column *b)* how many have at least one different value, and column *c)* and *d)* show how many have only values either in the English or in the French DBpedia, respectively (values in percentages are reported between brackets).

3 Extending the existing alignment

A large portion of the data extracted by the DBpedia community comes from the templates that are used in Wikipedia articles for synthetic descriptions. Templates define a set of attributes to describe a certain kind of entity (e.g. Authors, Football players, Cars, Planets). The task of mapping templates consists in matching attributes of a given template to properties of the DBpedia ontology. The DBpedia ontology is relatively large (more than 1500 properties for DBpedia - version 3.9) and manually finding the appropriate property to be mapped can take some time. However, many attributes are used with the same meaning in several templates, for instance *name*, *birth date* or *nationality* in templates for persons description. Avoiding the need to repeat these mappings would save DBpedia contributors a lot of time, and would speed up the mapping process.

We propose therefore an approach to expand the property mappings to all non ambiguous attributes, that is attributes that have always been manually mapped to the same ontology property. This results in the extension of the alignments between the properties textually generated from the attributes, and the ontology properties. And so, it extends the alignment between language specific datasets. By non-ambiguous attributes, we mean the terms that have not proven to be ambiguous in the existing mappings. The integration of the extended mappings into the mapping data would require human validation in order to check for incorrect alignments. In the following, we evaluate the possible gain obtained from the approach we propose. We use a simple heuristic to select mappings that are likely to be correctly propagated: we select only the attributes that have been mapped consistently to the same ontology property multiple times.

Concerning the mapping frequency of non ambiguous attributes in French DBpedia to the DBpedia ontology properties, 47 attributes are mapped at least twice, 18 attributes mapped at least three times (i.e. *lieu de décès* → `dbo:deathPlace`), and only one mapped at least ten times (i.e. *nom* → `foaf:name`). Since we assume that the mapping frequency is a good indicator of the correctness of the mapping,

in the rest of the section we will consider only the mappings that were mapped at least twice (i.e. frequency $\geq$2). Moreover, we carry out a manual validation of the 47 mappings appearing more than twice, to check if they are correct according to the attribute names. The results of such evaluation confirm that in 83% of the cases (i.e. 35 mappings), the mappings are correct. The validity of the remaining ones can be biased by the context in which they appear, since the attribute terms are either vague, or polysemous (i.e. could have different meanings). For instance, mapping the attribute *division* to `dbo:locatedInArea` seems correct for geographic places but *division* could be used to indicate also a football league or an organization department, and in those cases the mapping is incorrect.

generic prop. (p) `propFr:`	ontology prop. (po) `dbo:`	values for p	val. for p in po range (%)	values for po	values for both	same values (%)
`lieuDeDécès`	`deathPlace`	25477	14615 (57.3)	17190	13314	7579 (56.9)
`région`	`region`	87917	79853 (90)	51713	46077	45993 (99)
`nationalité`	`nationality`	44345	10071 (22.7)	46985	34884	8887 (25.4)
`lieuDeNaiss.`	`birthPlace`	66262	37326 (56.3)	49430	41716	24569 (58.8)
total object prop		645719	391044 (60.5)	482444	284201	209692 (73.7)
total datatype prop		680481	111876 (16.4)	517368	259623	59047 (22.7)
total		1326200	502920 (37.9)	999812	543824	268739 (49.4)

Table 2 Comparison of values between generic and ontology properties for the extended mappings in French DBpedia. Column *values for p* reports the number of instances that have values for the generic properties; column *val. for p in po range* reports the instances for which the generic property values are coherent with the ontology property signature; column *values for po* the instances that have values for the mapped ontology property. Column *values for both* reports the number of instances that have values for both the generic and the ontology property; column *same value* reports those for which the generic property and the ontology property have the same value.

Table 2 provides for each mapping a comparison between the number of instances that have a value for the generic property (build from the attribute occurence), and the number of instances that have a value for the mapped ontology property. For instance, the property `propFr:lieuDeDécès` is present for more than 25,000 instances (column *values for p*, Table 2), and `dbo:deathPlace` for more than 17,000 (column *values for po*). Note that *lieu de décès* is not the only attribute to be mapped to `dbo:deathPlace` (i.e. also *lieu décès*, *décès*, and other variants). The column *values for both* indicates how often the mapping *lieu de décès* to `dbo:deathPlace` is actually applied, and it gives the number of instances that have values for both the generic and the ontology property (i.e. 13,314). The potential gain of this mapping extension is given by the number of instances that have a value for the generic property, but no values for the ontology property, i.e. 25,477 - 13,314 = 12,163 additional values for `dbo:deathPlace`. Over the 47 mappings that can be extended, the potential gain is 1,326,200 - 543,824 = 782,376, corresponding to an increase of about 59%.

Column *same values* gives the number of instances for which the generic property and the ontology property have the same value. However, the comparison with the number of co-occurrence of the two properties is not fair, as the extractor that generates the values for the ontology property is guided by the property signature (in the example of `dbo:deathPlace`, the expected value is an instance), whereas the generic property is more subject to noise and may generate another output from the same attribute value (for instance a number if the attribute value begins with a street number). So for this comparison, we narrow our scope to the instances for which the generic property values are coherent with the ontology property signature (column *values for p in po range*). Out of the 25,477 instances that have a value for `propFr:lieuDeDécès`, only 14,615 have an object value. However, every time there is an object value for `propFr:lieuDeDécès` and a value for `dbo:deathPlace`, these are the same. In a symmetric way, we calculated the mapping frequency of non ambiguous attributes in English DBpedia to ontology properties. As expected, many more attributes are mapped more frequently than in the French chapter (i.e. 689 attributes are mapped at least twice, 296 attributes mapped at least five times, and 160 mapped at least ten times, e.g. *twin* to `dbo:twinCity` or *successor* to `dbo:successor`).

generic property `propFr:`	ontology property `dbo:`	new values wrt. DBpedia En and Fr	same values (%)	diff. values
`lieuDeDécès`	`deathPlace`	4393	4016 (89.3)	479
`région`	`region`	16491	18496 (82.5)	3906
`nationalité`	`nationality`	358	870 (81.3)	200
`lieuDeNaissance`	`birthPlace`	6934	7016 (89)	862
total object prop		85951	73306 (82.7)	15250
total datatype prop		16155	45177 (90)	5001
total		102106	118483 (85.4)	20251

Table 3 Comparison between values obtained with the mappings extension in French and English DBpedia. Column *new values wrt. DBpedia En and Fr* provides the number of values that were added through the mapping extension process w.r.t. the values already available through ontology properties in English and French DBpedia.

To evaluate the quality of the data obtained applying the above presented approach to extend the mapping among language specific DBpedia chapters, we compare the values obtained from the mappings extension for the French chapter, with the values obtained for the English chapter, as previously done in Section 2 for the existing alignments. Table 3 summarizes the results obtained from such comparison. More specifically, it provides the number of values that were added through this process (column *new values wrt. DBpedia En and Fr*) with respect to the values already available through ontology properties in English and French DBpedia. For instance, the mapping extension (*lieu de naissance* to `dbo:birthPlace`) considered earlier generates 6,934 new values. Among the values that were already present in the English chapter, 7,016 are the same and 862 differ (89% identical). We can notice that this is about the same ratio as for the comparison between values for the

same ontology property in Section 2, i.e. 14,139 identical values (column a, Table 1) and 1,965 different (columns a+b, Table 1), i.e. 87% identical. We can consider it as a positive result, as it suggests that most of the differences in the values are generated by differences between the two chapters of DBpedia, rather than from mappings mistakes.

Concerning the 47 mappings described in Section 3, we have 118,483 identical values and 20,251 different values (respectively, columns *same values* and *different values* in Table 3). If we consider object properties and datatype properties separately, we obtain now a better correlation between values of English and French chapters for datatype properties (90% instances with same values) than for object properties (82%). This may be explained by the fact that many datatypes are not specified for generic properties (e.g. for strings), so we selected the values that fit the range of the property as specified in the ontology, and we removed values that generated noise in the comparison described in Section 2.

4 QA experimental setting

To benefit from the amount of information coming from the aligned language specific datasets described before, we extended QAKiS, our system for open domain Question Answering over linked data (Cabrio et al., 2012), to query language specific DBpedia chapters (Section 4.1). To enhance users interactions with the Web of Data, query interfaces providing a flexible mapping between natural language expressions, and concepts and relations in structured knowledge bases are becoming particularly relevant. More specifically, QAKiS allows end users to submit a query to an RDF triple store in English and obtain the answer in the same language, hiding the complexity of the non intuitive formal query languages involved in the resolution process. At the same time, the expressiveness of these standards is exploited to scale to the huge amounts of available semantic data. We evaluate QAKiS extension to query English, French and German DBpedia chapters with two sets of experiments, described in Section 4.2 and 4.3.

4.1 QA system description: QAKiS

QAKiS (Question Answering wiKiFramework-based System)[6] (Cabrio et al., 2012) addresses the task of QA over structured knowledge-bases (e.g. DBpedia), where the relevant information is expressed also in unstructured forms (e.g. Wikipedia pages). It implements a relation-based matching for question interpretation, to convert the user question into a query language (e.g. SPARQL). More specifically, it makes use of relational patterns - automatically extracted from Wikipedia and collected in

[6] A demo is available at `http://qakis.org/qakis2/`

the WikiFramework repository (Mahendra, Wanzare, Bernardi, Lavelli, & Magnini, 2011) - that capture different ways to express a certain relation in a given language.[7] QAKiS is composed of four main modules (Fig. 2):

- the *query generator* takes the user question as input, generates the typed questions, and then generates the SPARQL queries from the retrieved patterns;
- the *pattern matcher* takes as input a typed question and retrieves the patterns (among those in the repository) matching it with the highest similarity;
- the *SPARQL package* handles the queries to DBpedia; and
- a *Named Entity (NE) Recognizer*.

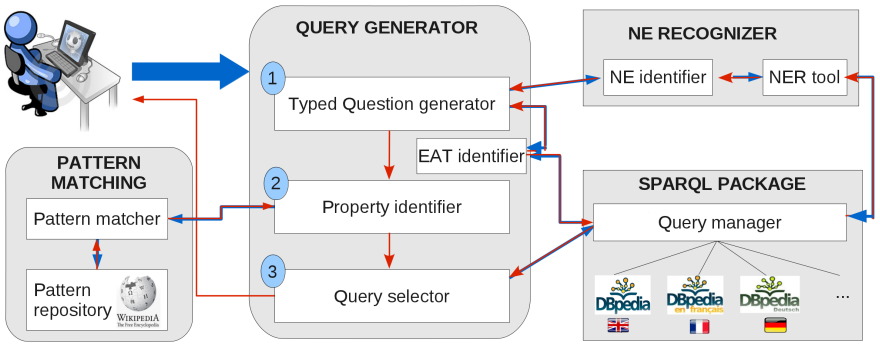

Fig. 2 QAKiS workflow

The actual version of QAKiS targets questions containing a NE related to the answer through one property of the ontology, as *Which river does the Brooklyn Bridge cross?*, or *In which country does the Nile starts?*. Such questions match a single pattern (i.e. one relation).

Before running the *pattern matcher* component, the target of the question is identified using the Stanford Core NLP NE Recognizer[8], together with a set of strategies based on the comparison with the labels of the instances in the DBpedia ontology. Then a *typed question* is generated by replacing the question keywords (e.g. who, where) and the NE by their types and supertypes. A Word Overlap algorithm is then applied to match such typed questions with the patterns for each relation. A similarity score is provided for each match: the highest represents the most likely relation. A set of patterns is retrieved by the pattern matcher component for each typed question, and sorted by decreasing matching score. For each of them, a set of SPARQL queries is generated and then sent to the SPARQL endpoint for answer retrieval.

[7] Gerber et al. [this volume] describe another framework (i.e. BOA) to address the challenge of extracting structured data as RDF from unstructured data.

[8] `http://nlp.stanford.edu/software/CRF-NER.shtml`

4.1.1 QAKiS extension to query language specific DBpedia chapters

To allow QAKiS to query the ontology properties of language specific DBpedia chapters, we modified QAKiS architecture at the *SPARQL package* level. The typed questions generation and the pattern matching steps work as before, but now, instead of sending the query to English DBpedia only, the *query manager* reformulates the query and sends it to multiple DBpedia chapters. As only the English chapter contains labels in English, this change has no impact on the NE Recognition. The main difference lies in the query selection step. As before, patterns are considered in order of decreasing matching score, the generated query is then evaluated and if no results are found the next pattern is considered, and so on. However, as queries are now evaluated on several DBpedia chapters, it is more likely to get results, terminating query selection with a higher matching score. Currently, the results of a SPARQL query are aggregated by the set union. Other strategies could be considered, such as using a voting mechanism to select the most frequent answer, or enforcing a priority according to data provenance (e.g. English chapter could be considered as more reliable for questions related to English culture). In the current version, QAKiS allows to query English, French and German DBpedia chapters.

4.2 Evaluation on QALD-2 data set

As a first step of our experiments, we evaluate if the integration of the French and German DBpedia datasets has an impact on QAKiS performances on the standard benchmark of the QALD-2 challenge[3] (DBpedia track). It is provided by QALD organizers to compare different approaches and systems that mediate between a user, expressing his or her information need in natural language, and semantic data. Since in the actual version of the system it targets only questions containing a NE related to the answer through one property of the ontology (e.g. *In which military conflicts did Lawrence of Arabia participate?*), we extracted from the complete benchmark the questions corresponding to such criteria. Out of 100 questions available for testing, the questions containing a NE related to the answer through one property of the ontology amount to 32, which we used in our experiment. The discarded questions require either some forms of reasoning (e.g. counting or ordering) on data, aggregation (from datasets different from DBpedia), involve n-ary relations, or they are boolean questions. We run both QAKiS$_{EN}$ (i.e. the system taking part into the challenge) and QAKiS$_{EN+FR}$ and QAKiS$_{EN+DE}$ (the versions enriched with the French and German DBpedia chapters, respectively) on the reduced set of questions.

Since the answer to QALD-2 questions can be retrieved from the English DBpedia, we do not expect multilingual QAKiS to improve its performances. On the contrary, we want to verify that QAKiS performances do not decrease (due to the choice of the wrong relation triggered by a different pattern that finds an answer in language specific DBpedia chapters). QAKiS$_{EN}$ correctly answers to 15/32 questions and partially correctly to 4/32 questions (e.g. in *Give me all companies in*

Munich the list provided by QAKiS using `foundationPlace` as relation and *Munich* as subject, is only partially overlapping with the one proposed by the organizers). The extended QAKiS system often selects patterns that are different with respect to the one selected by the original system, but except in one case the identified target relation is the same, meaning that performances are not worsen when querying several language specific DBpedia chapters.

4.3 Separate evaluations on French and German DBpedia chapters

As introduced before, the questions created for QALD-2 challenge are thought to find an answer in the English DBpedia, so they cannot be used to evaluate the contribution resulting from the extension of property alignments to language specific DBpedia chapters. Since we are not aware of any standard list of questions whose answers can be found in French and German DBpedia chapters only, we created our reference set to evaluate the extension in QAKiS$_{EN+FR}$ and QAKiS$_{EN+DE}$'s coverage performing the following steps:

1. we take the sample of 32 questions from QALD-2;
2. we extract the list of triples present in French (and German) DBpedia only (as described in Section 2);
3. in each question we substitute the named entity with another entity for which the asked relation can be found in the French (or German) chapter only.

For instance, for QALD-2 question: *How tall is Michael Jordan?*, we substitute the Named Entity *Michael Jordan* with the entity *Margaret Simpson*, for which we know that the relation `height` is not present in English DBpedia, but it is linked in the French chapter. As a result, we obtain the question *How tall is Margaret Simpson?*, that we submit to QAKiS$_{EN+FR}$. Following the same procedure for German, in *Who developed Skype?* we substituted the NE *Skype* with the entity *IronPython*, obtaining the question *Who developed IronPython?*.[9] For some properties (i.e. `Governor`, `Battle`, `FoundationPlace`, `Mission` and `RestingPlace`), no additional links are provided by language specific DBpedia chapters, so we discarded related questions.

QAKiS precision on the new set of questions over French and German DBpedia is in line with QAKiS$_{EN}$ on English DBpedia ($\sim$ 50%) (i.e. out of 27 questions, QAKiS$_{EN+FR}$ correctly answers to 14 questions and partially correctly to 1 question). To double-check, we run the same set of questions on QAKiS$_{EN}$ (that relies on the English chapter only), and in no cases it was able to detect the correct answer, as expected. This second evaluation did not have the goal to show improved performances of the extended QAKiS system with respect to its precision, but to show that the integration of language specific DBpedia chapters in the system is easily achievable, and that the expected improvements on its coverage are really promising and worth exploring (see Table 1).

[9] The obtained set of transformed questions is available at `http://qakis.org/qakis2/`

5 Related work

In this chapter, we have exploited existing alignments over DBpedia data to compare and aggregate data from different Wikipedia chapters. The instance alignments are manually edited by the Wikipedia community (as interlanguage links), and the property alignments by the DBpedia community. The field of ontology alignment tackles questions about automated or partially automated alignments techniques. Rahm and Bernstein (2001), Shvaiko and Euzenat (2013) present surveys on the topic.

Several works address the more specific topic of data integration from Wikipedia chapters directly from the article content. Rinser et al. (2013) provide an overview of instance-based template-attributes matching approaches over language specific Wikipedia chapters. They also present their own, very thorough approach. First, several criteria are taken into account to improve the instance matching resulting from the inter-language links (i.e. based on this instance alignment, a template alignment is computed according to their use in matched instances). Then, attributes of aligned templates are matched according to the instances and values they relate.

To predict the matching probability of pairs of infobox attributes instances across different language versions, Adar, Skinner, and Weld (2009) employ self-supervised machine learning with a logistic regression classifier using a broad range of features (e.g. n-gram/word overlap of attribute keys and values, wiki link overlap). Bouma, Duarte, and Islam (2009) perform a matching of infobox attribute based on instance data. Bouma (2010) describes a system for linking the thesaurus of the Netherlands Institute for Sound and Vision to EnglishWordNet and DBpedia, using EuroWordNet and Dutch Wikipedia as intermediaries for the two alignments. Tacchini, Schultz, and Bizer (2009) provide several strategies for merging data extracted from different Wikipedia chapters. They present a software framework for fusing RDF datasets based on different conflict resolution strategies, and they apply it to fuse infobox data that is extracted from multilingual editions of Wikipedia.

Aprosio, Giuliano, and Lavelli (2013) define a methodology to increase DBpedia coverage in different languages. Information is bootstrapped through cross-language links, starting from the available mappings in some pivot languages, and then extending the existing DBpedia datasets comparing the classifications in different languages. When such classification is missing, supervised classifiers are trained on the original DBpedia (relying on the Distant Supervision paradigm).

A survey on the field of Question Answering is provided by (Lopez, Uren, Sabou, & Motta, 2011), with a focus on ontology-based QA. Moreover, they examine the potential of open user friendly interfaces for the SW to support end users in reusing and querying SW content. State of the art QA systems over Linked Data generally address the issue of question interpretation mapping a natural language question to a triple-based representation. For instance, Freya (Damljanovic, Agatonovic, & Cunningham, 2012) uses syntactic parsing in combination with ontology-based lookup for question interpretation, partly relying on the user's help in selecting the entity that is most appropriate as match for some natural language expressions. One of the problems of that approach is that often end-users are unable to help, in case they are not informed about the modeling and vocabulary of the data. PowerAqua (Lopez,

Uren, Sabou, & Motta, 2009) accepts user queries expressed in NL and retrieves answers from multiple semantic sources on the SW. It follows a pipeline architecture, according to which the question is *i)* transformed by the linguistic component into a triple based intermediate format, *ii)* passed to a set of components to identify potentially suitable semantic entities in various ontologies, and then *iii)* the various interpretations produced in different ontologies are merged and ranked for answer retrieval. PowerAqua's main limitation is in its linguistic coverage.
Pythia (Unger & Cimiano, 2011) relies on a deep linguistic analysis to compositionally construct meaning representations using a vocabulary aligned to the vocabulary of a given ontology. Pythia's major drawback is that it requires a lexicon, which has to be manually created. More recently, an approach based on Pythia (Unger & Cimiano, 2011) but more similar to the one adopted in QAKiS is presented (Unger et al., 2012). It relies on a linguistic parse of the question to produce a SPARQL template that directly mirrors the internal structure of the question (i.e. SPARQL templates with slots to be filled with URIs). This template is then instantiated using statistical entity identification and predicate detection.

6 Conclusions and future work

In the first part of this chapter we have proposed an in-depth comparative analysis of language specific DBpedia chapters, focusing in particular on the French and the English DBpedia chapters, proving that most of their content is complementary: each chapter brings a significant amount of data that cannot be found in the other chapter (about half of the data from the French DBpedia and 80% of the data from the English DBpedia). To perform this comparison, we have first considered the existing alignments and compared the two chapters to highlight their differences. Then, we have proposed an approach to extend the existing property alignments to all the occurrences of non-ambiguous attributes (i.e. attributes that humans have always mapped to the same ontology properties).

Since the DBpedia ontology is continuously evolving, maintaining its consistency is a complex task that needs continual updates. Some studies have been carried out to evaluate the quality of the DBpedia ontology: being able to automatically compare the values of several chapters, as we showed in our work, could provide interesting indicators of errors or vandalism in one chapter, and detect discrepancies among vocabulary used among chapters, or even among topics of the same chapter.

In the second part of this chapter, we have considered Question Answering over Linked Data scenario. To show the interesting potential for NLP applications resulting from the property alignments in language specific DBpedia chapters, we have extended the QAKiS system so that it can query the ontology properties of the French and German DBpedia chapters. We show that this integration extends the system coverage (i.e. the recall), without having a negative impact on its precision.

We plan to extend the presented work in a number of directions. First, we would like to improve the mapping extension approach by taking into account instance

types to disambiguate attributes. We also plan to use alignment tools (e.g. Silk[10]) to suggest additional property alignments based on the similarity of their use in their respective chapters (e.g. considering the number of equivalent pairs that two properties have in common). Moreover, since the pieces of information obtained by querying distributed SPARQL endpoints may provide different results for the same query, leading to an inconsistent set of information about the same topic, we are investigating the problem of reconciling information obtained by distributed SPARQL endpoints. In particular, we plan to address this problem by combining the AI nonmonotonic reasoning framework called argumentation theory to reason over inconsistent sets of information, and provide nevertheless a unique and motivated answer to the user. We are currently working at the implementation and evaluation of such a framework in QAKiS (Cabrio, Cojan, Villata, & Gandon, 2013).

References

Adar, E., Skinner, M., & Weld, D. S. (2009). Information arbitrage across multilingual Wikipedia. In *Proceedings of the second acm international conference on web search and data mining (WSDM)* (pp. 94–103). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/1498759.1498813 doi: 10.1145/1498759.1498813

Aprosio, A. P., Giuliano, C., & Lavelli, A. (2013). Towards an Automatic Creation of Localized Versions of DBpedia. In *Proceedings of the international semantic web conference (ISWC)* (p. 494-509).

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., & Hellmann, S. (2009). DBpedia - A crystallization point for the Web of Data. *Web Semant.*, *7*(3), 154–165. Retrieved from http://dx.doi.org/10.1016/j.websem.2009.07.002 doi: 10.1016/j.websem.2009.07.002

Bouma, G. (2010). Cross-lingual Ontology Alignment using EuroWordNet and Wikipedia. In *Proceedings of the language resources and evaluation conference (LREC)*.

Bouma, G., Duarte, S., & Islam, Z. (2009). Cross-lingual alignment and completion of Wikipedia templates. In *Proceedings of the third international workshop on cross lingual information access: Addressing the information need of multilingual societies (CLIAWS3)* (pp. 21–29). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=1572433.1572437

Buitelaar, P., Choi, K.-S., Cimiano, P., & Hovy, E. H. (2013). The Multilingual Semantic Web (Dagstuhl Seminar 12362). *Dagstuhl Reports*, *2*(9), 15–94. Retrieved from http://drops.dagstuhl.de/opus/volltexte/2013/3788 doi: http://dx.doi.org/10.4230/DagRep.2.9.15

[10] http://lod2.eu/Project/Silk.html

Cabrio, E., Cojan, J., Palmero, A., Magnini, B., Lavelli, A., & Gandon, F. (2012). QAKiS: an Open Domain QA System based on Relational Patterns. In *Proceedings of the international semantic web conference (ISWC demos).*

Cabrio, E., Cojan, J., Villata, S., & Gandon, F. (2013). Hunting for Inconsistencies in Multilingual DBpedia with QAKiS. In *Proceedings of the international semantic web conference (ISWC posters & demos)* (p. 69-72).

Damljanovic, D., Agatonovic, M., & Cunningham, H. (2012). FREyA: an Interactive Way of Querying Linked Data using Natural Language. In *Proceedings of the 8th international conference on the semantic web (ESWC)* (pp. 125–138). Springer-Verlag. Retrieved from http://dx.doi.org/10.1007/978-3-642-25953-1_11 doi: 10.1007/978-3-642-25953-1_11

Kontokostas, D., Bratsas, C., Auer, S., Hellmann, S., Antoniou, I., & Metakides, G. (2012). Internationalization of Linked Data: The case of the Greek DBpedia edition. *Web Semantics: Science, Services and Agents on the World Wide Web*, *15*(0), 51 - 61. Retrieved from http://www.sciencedirect.com/science/article/pii/S1570826812000030 doi: 10.1016/j.websem.2012.01.001

Lopez, V., Uren, V. S., Sabou, M., & Motta, E. (2009). Cross Ontology Query Answering on the Semantic Web: an Initial Evaluation. In *Proceedings of the 5th international conference on knowledge capture (K-CAP)* (p. 17-24).

Lopez, V., Uren, V. S., Sabou, M., & Motta, E. (2011). Is question answering fit for the semantic web?: A survey. *Semantic Web*, *2*(2), 125-155.

Mahendra, R., Wanzare, L., Bernardi, R., Lavelli, A., & Magnini, B. (2011). Acquiring Relational Patterns from Wikipedia: A Case Study. In *Proceedings of the 5th language and technology conference (LTC).*

Rahm, E., & Bernstein, P. A. (2001). A Survey of Approaches to Automatic Schema Matching. *The VLDB Journal*, *10*(4), 334–350. Retrieved from http://dx.doi.org/10.1007/s007780100057 doi: 10.1007/s007780100057

Rinser, D., Lange, D., & Naumann, F. (2013). Cross-lingual Entity Matching and Infobox Alignment in Wikipedia. *Inf. Syst.*, *38*(6), 887-907.

Shvaiko, P., & Euzenat, J. (2013). Ontology Matching: State of the Art and Future Challenges. *IEEE Trans. Knowl. Data Eng.*, *25*(1), 158-176.

Tacchini, E., Schultz, A., & Bizer, C. (2009). Experiments with Wikipedia Cross-Language Data Fusion. In S. Auer, C. Bizer, & G. A. Grimnes (Eds.), (Vol. 449). Retrieved from http://CEUR-WS.org/Vol-449/Paper3.pdf

Unger, C., Bühmann, L., Lehmann, J., Ngonga Ngomo, A.-C., Gerber, D., & Cimiano, P. (2012). Template-based Question Answering over RDF Data. In *Proceedings of the 21st international conference on world wide web (WWW)* (pp. 639–648). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2187836.2187923 doi: 10.1145/2187836.2187923

Unger, C., & Cimiano, P. (2011). Pythia: Compositional Meaning Construction for Ontology-Based Question Answering on the Semantic Web. In *Proceedings of the 16th international conference on applications of natural language to information systems (NLDB)* (p. 153-160).

Index

DBpedia, 2

language specific DBpedia chapters, 4

Multilingual Question Answering over Linked Data challenge, 3
Multilingual Semantic Web, 2

property alignment, 4

QAKiS, 10
Question Answering over Linked Data, 3

relational patterns, 10

users interactions, 10

Web of Data, 10
WikiFramework, 10
Wikipedia, 2